

Remarks

Applicants thank the Examiner for his careful consideration of their Application.

Reconsideration of this Application is now respectfully requested.

Upon entry of the foregoing Amendment, Claims 1-43 are pending in the application, with Claims 1, 29, 30, 33, 34, 37, 38, and 42 being the independent claims. New Claim 43 has been added.

Based on the above Amendment and the following Remarks, Applicants respectfully request that the Examiner reconsider all outstanding objections and rejections and that they be withdrawn.

Objection to the Specification

At Page 1, the Office Action objects to the specification as containing “embedded hyperlinks and/or other form [*sic*] of browser-executable code.” Applicants have reviewed the specification and have amended it to remove the one hyperlink found (at Page 1, line 14).

Rejections under 35 U.S.C. § 112

At Page 1, the Office Action rejects Claim 10 under 35 U.S.C. § 112, second paragraph, as being indefinite, based its containing the term “very.” In response, Applicants have amended Claim 10 such that it no longer contains “very.” Support for this amendment may be found in the specification at, for example, Page 12, lines 7-21, Fig. 3, Page 18, line 13 to Page 19, line 5, and Fig. 4.

Rejections under 35 U.S.C. § 102 and under 35 U.S.C. § 103

The Office Action, at Pages 1-7, rejects Claims 1-6, 8, 10-14, 19-21, and 23-42 under 35 U.S.C. § 102(e) as being anticipated by Aiken (U.S. Patent No. 6,240,409). At Pages 7-8, the Office Action rejects Claims 7, 9, 15-18, and 22 under 35 U.S.C. § 103(a) as being unpatentable over Aiken. Applicants respectfully traverse these rejections for the following reasons.

The invention, for example, as claimed in Claims 1 and 29, is directed toward the detection of similar documents. A document is obtained and filtered, and a document identifier and hash value are determined for the document. The document identifier and hash value are used to generate a tuple corresponding to the filtered document. The tuple is compared to other tuples in a document storage structure. Similarity to another document is determined if the tuple is clustered with another tuple in the document storage structure.

In contrast, Aiken, while disclosing the detection of document similarity, discloses a different way of performing such detection. One general difference between the claimed invention and Aiken lies in that the focus of Aiken is on syntax and sub-string analysis, while the claimed invention focuses more on semantic processing (for example, the removal of “unimportant words,” discussed at the bottom of Col. 4, is a syntactic technique, as opposed to the filtering disclosed in Applicants’ disclosure, which is a semantic technique). A further general difference is that the claimed invention relies on information retrieval techniques, while Aiken appears not to do so. These, and other differences, are reflected in the further discussion below.

As discussed above, Claims 1 and 29 include the determination of “a hash value **for the filtered document.**” In contrast, Aiken computes hash values **for sub-strings**, as discussed, for example, at Col. 5, lines 10-35. That is, while the claimed invention determines hash values on a document basis, Aiken computes hash values on a sub-string basis. Therefore, Aiken does **not** disclose determination of “a hash value for the filtered document.”

Similarly, Claims 1 and 29 include generating “a tuple **for the filtered document.**” The claimed tuple comprises the **document identifier for the filtered document** and the hash value **for the filtered document.** In contrast, Aiken, noting, for example, Col. 4, line 65 to Col. 5, line 4, forms pairs consisting of a hash value **for each sub-string** and a position **of the substring within the document.** That is, Aiken’s pairings are formed on a sub-string basis, not on a document basis. Therefore, Aiken does **not** disclose generating “a tuple for the filtered document.”

Furthermore, Claims 1 and 29 include comparing a tuple generated **for a document** with a plurality of tuples **representing other documents.** In contrast, the <hash value, position> pairs formed in Aiken are compared with other <hash value, position> pairs similarly formed, and thus **corresponding to sub-strings, not documents.** Therefore, Aiken does **not** disclose comparing a tuple with a plurality of other tuples, each representing one of a plurality of documents.

It is respectfully submitted that, for at least these reasons, Claims 1 and 29 are allowable over the cited prior art. Hence, it is further submitted that Claims 2-28, which depend from Claim 1, are also allowable over the cited prior art.

Claims 30 and 33 contain the same limitations discussed in the arguments with respect to Claims 1 and 29; hence, those arguments apply equally to Claims 30 and 33. A number of additional arguments are also applicable.

First, the invention claimed in Claims 30 and 33 includes **retaining only retained tokens using at least one token threshold**. The Office Action, noting the discussion of Claim 3 at Page 4, cites Aiken at Col. 11, lines 15-30 as disclosing this limitation. However, Col. 11, lines 15-30 address **the retention of documents by comparing a match ratio for a document with a (document) threshold**. While the match ratio is determined by comparing the hash values (of the sub-strings) of two documents and determining the ratio of the number of hash values in common to the total number of hash values in one of the documents (see, e.g., Aiken at Col. 11, lines 1-9), **the threshold comparison is not performed on a token basis, but rather on a document basis**.

Second, the invention claimed in Claims 30 and 33 includes **arranging retained tokens into an arranged token stream**. Also, in Claims 30 and 33, **subsequent processing**, beginning with obtaining a hash value for a document, **is performed on the tokens of the arranged token stream**. The Office Action, noting the discussion of Claim 4 at Page 4, relies on Fig. 4a, step 404 of Aiken for disclosure of such arranging; however, it is respectfully submitted that this can not correspond to the claimed arrangement of tokens. First, noting Col. 10, lines 49-54, step 404 **sorts pairs, which**, as discussed above, **comprise a hash value and a position**, based on the position such that position values from a previously-indexed document are grouped together

sequentially. **The pairs are not the tokens**, as claimed. Second, step 404 can not correspond to this arranging of tokens because **the tokens in Claims 30 and 33 lack position values on which step 404 is based.**

Finally, the invention as claimed in Claims 30 and 33 includes inserting a tuple for a document into a document storage tree and determining that the document is similar to another document if the tuple, after being inserted, is collocated with a tuple corresponding to another document. The Office Action, noting the discussion of Claim 26 at Pages 5-6, relies on Fig. 4c and Col. 8, lines 31-54 of Aiken for disclosure of these features. Applicants respectfully submit, however, that these portions of Aiken do not disclose the claimed features. Col. 8, lines 31-54 of Aiken addresses a data structure that stores names of documents, ranges of bytes in the corresponding documents, and, optionally, total numbers of hashes in the respective documents. As explained at Col. 8, lines 48-54, the <hash value, position> pairs of Aiken are **not inserted** into this data structure, but rather, this data structure is used to retrieve data for a pair based on the position. Col. 8, lines 31-54 does not address anything resembling the determination of similarity.

Fig. 4c and its accompanying explanation at Col. 11, line 47 to Col. 12, line 2 of Aiken address clustering of a current document with existing clusters of documents using a union-find algorithm. However, this algorithm does not address collocated tuples, as claimed, in order to determine similarity of documents. Rather, the union-find algorithm of Aiken determines a set to which a document belongs and associates it with (merges it into) that set.

For at least these further reasons, it is respectfully submitted that Claims 30 and 33 are allowable over the cited prior art. Hence, it is further submitted that Claims 31 and 32, which depend from Claim 30, are also allowable over the cited prior art.

Claims 34 and 37 recite determining a hash value **for a document**, accessing a document storage structure comprising hash values representing a plurality of documents, and detecting if a document is similar to another document in the storage structure based on determining if their hash values are equivalent. Given that the claimed hash values are computed on a document-by-document basis, the same arguments applied to Claims 1 and 29 are equally applicable here. Also, the Office Action relies on Aiken, noting Fig. 4c and Col. 11, line 47 to Col. 12, line 2 for a disclosure of determining equivalent hash values. However, nowhere in the figure or in the cited text is there any mention of determining equivalence of hash values as a determination of similarity.

For at least these reasons, it is respectfully submitted that Claims 34 and 37 are allowable over the cited prior art. Therefore, Claims 35 and 36, which depend from Claim 34, are further submitted as being allowable over the cited prior art.

Claims 38 and 42 recite comparing a document to documents in a document collection using a hash algorithm and collection statistics to detect if the document is similar to any of the documents in the document collection. The Office Action, in particular, cites Aiken, Fig. 4c and Col. 11, line 47 to Col. 12, line 2, as disclosing the use of collection statistics in clustering similar documents. However, **there is no mention of any collection statistics in the cited**

figure or passage from Aiken. What is discussed is the use of a union-find algorithm. As discussed above, a union-find algorithm determines a set to which a document belongs and associates it with (i.e., merges it into) that set. However, there is no disclosure in Aiken of a union-find algorithm using collection statistics.

Also, noting Col. 11, lines 56-60, the comparison performed in the clustering of Aiken is a comparison **between the current document and a single document** (from an existing cluster), **meaning that collection statistics would not be used**, as in the claimed invention.

It is further noted that, at Col. 2, lines 36-41 and 47 (the latter line establishing a connection between Aiken's disclosed invention and the desirable properties described in the former lines), **Aiken specifically teaches away from the use of "probability for measuring comparison accuracy."** As statistics are probabilistic in nature, this would teach away from the use of collection statistics.

For at least these reasons, it is respectfully submitted that Claims 38 and 42 are allowable over the cited prior art. It is further submitted that Claims 39-41, which depend from Claim 38, are thus also allowable over the cited prior art.

In view of the above, it is respectfully submitted that all of Claims 1-42 are allowable over the cited prior art in view of the allowability of all of the independent claims. Applicants additionally present the following arguments regarding various dependent claims:

- Claim 3 includes "retaining a token in the token stream as a retained token according to at least one token threshold." Claim 3 is thus further allowable over the cited prior

art in view of the argument discussed above in connection with Claim 30 in connection with a similar limitation, in addition to the arguments made in connection with Claim 1.

- Claim 4 includes “arranging the retained tokens in the token stream to obtain an arranged token stream.” Claim 4 is thus further allowable over the cited prior art in view of the argument discussed above in connection with Claim 30 in connection with a similar limitation, in addition to the arguments made in connection with Claims 1 and 3.
- Claim 5 includes “determining a hash value [for a filtered document] by individually processing each retained token in the token stream.” In addition to the fact that Aiken does not determine a hash value for a document (but, instead, determines hash values for sub-strings), Aiken, in the cited passages (Col. 6, lines 7-28 and Col. 9, lines 24-26), applies hashing to all sub-strings, not only to retained tokens. Claim 5 is thus allowable over the cited prior art for this further reason, in addition to the arguments made in connection with Claims 1 and 3.
- Claim 6 recites a step of “determining a score for each token in the token stream” and a step of “comparing the score for each token to a first token threshold.” Claim 7 depends from Claim 6 and recites a further step of “comparing the score for each retained token [i.e., in a further step of Claim 6 noted below] to a second token threshold.” These claims thus add limitations similar to the limitation added by

Claim 3, so the above arguments with respect to Claim 3 (and Claim 30), therefore, also apply to these claims. Furthermore, the Office Action cites Col. 11, lines 15-30 of Aiken as disclosing the limitations of Claim 6, including the step of modifying the token stream;" however, as the cited passage is concerned with comparing and discarding documents, not tokens, it does not disclose this limitation. Therefore, Claims 6 and 7 are allowable over the cited prior art for these further reasons, in addition to the arguments made in connection with Claim 1.

- Claim 10, as amended, now recites the step of "removing a token from the token stream based on collection statistics and at least one token threshold." As previously discussed, neither collection statistics (discussed in connection with Claim 38) nor token thresholds (discussed in connection with Claim 30) are used in Aiken. Hence, Claim 10 is allowable over the cited prior art for these further reasons, in addition to the arguments made in connection with Claim 1.
- Claims 13 and 14 are directed to the use of collection statistics for filtering documents. The arguments presented above in connection with a similar limitation in Claim 38 are thus applicable to these claims, providing a further basis (in addition to the arguments made in connection with Claim 1) for the allowability of Claims 13 and 14 over the cited prior art.
- Claim 14 recites that "the collection statistics [used in filtering the document] pertain to the plurality of documents." In contrast, noting Col. 11, lines 1-14, Aiken uses

pairwise comparisons between two documents, not collection statistics based on a plurality of documents. Claim 14 is, therefore, allowable over the cited prior art for this further reason, in addition to the reasons discussed above.

- Claim 39 recites that “the collection statistics [used in the comparing step of Claim 38] pertain to the document collection [discussed in Claim 38].” In particular, Claim 38, from which Claim 39 depends, recites “a plurality of documents in a document collection.” Therefore, the argument presented in connection with Claim 14, in the immediately preceding paragraph, also provides a further basis for the allowability of Claim 39 over the cited prior art, in addition to the arguments made in connection with Claim 38.
- Claims 25 and 26 are directed to storage structures and to the determination of similarity of documents whose tuples are collocated in the storage structures. The arguments presented in connection with a similar limitation in Claim 30 are, therefore, also applicable to these claims, thus providing a further basis for allowability of Claims 25 and 26 over the cited prior art, in addition to the arguments made in connection with Claim 1.

Added Claim

New Claim 43 has been added as a dependent claim from Claim 1 and is, therefore, allowable, at least, as depending from an allowable claim. Note that the filtering techniques disclosed by Applicants in Figs. 2 and 3 and at Page 11-15 fall into the category of semantic

Applicants: FRIEDER et al.
Appl. No. 09/629,175

filtering techniques (see, particularly, Page 12, lines 15-16 and the definition of "similar" at Page 9, lines 11-15).

Applicants: FRIEDER et al.
Appl. No. 09/629,175

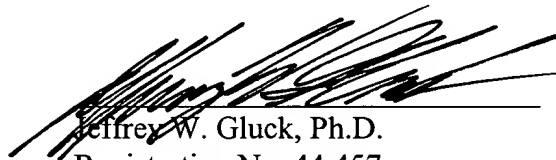
Conclusion

All of the stated grounds of objection and rejection have been properly traversed, accommodated, or rendered moot. Applicants, therefore, respectfully request that the Examiner reconsider all presently outstanding objections and rejections and that they be withdrawn. Applicants believe that a full and complete reply has been made to the outstanding Office Action and, as such, the present application is in condition for allowance. If the Examiner believes, for any reason, that personal communication will expedite prosecution of this application, the Examiner is hereby invited to telephone the undersigned at the number provided.

Prompt and favorable consideration of this Amendment is respectfully requested.

Respectfully submitted,

Date: December 9, 2002



Jeffrey W. Gluck, Ph.D.

Registration No. 44,457

VENABLE

P.O. Box 34385

Washington, D.C. 20043-9998

Telephone: (202) 962-4800

Direct Dial: (202) 216-8017

Telefax: (202) 962-8300

**APPENDIX: MARKED-UP VERSIONS OF AMENDED PORTIONS OF THE
APPLICATION**

In the Specification:

The paragraph beginning at Page 1, line 14 has been rewritten as follows:

[2] ~~http://nccam.nih.gov~~, The National Institutes of Health (NIH), National Center for Complementary and Alternative Medicine (NCCAM), April 12, 2000.

The paragraph beginning at Page 15, line 25 has been rewritten as follows:

Preferably, the hash value for the filtered document is determined using a hash algorithm having an approximately even distribution of hash values. More preferably, the hash value for the filtered document is determined using a secure hash algorithm. With a secure hash algorithm, the probability of two token streams creating the same hash value is reduced. Even more preferably, the hash value for the filtered document is determined using the hash algorithm SHA-1. With the hash algorithm SHA-1, each retained token in the arranged token stream is processed individually to obtain a hash value for the filtered document. With the SHA-1 hash algorithm, which uses 160 bits, the probability of duplicate values for different token streams is $P(2^{-160}) \underline{O(2^{-160})}$.

In the Claims:

Claim 10 has been amended as follows:

Applicants: FRIEDER et al.
Appl. No. 09/629,175

10. (Amended) A method as in claim 2, wherein the step of filtering further comprises removing a token from the token stream ~~if the token is either a very frequent token or a very infrequent token~~ based on collection statistics and at least one token threshold.

New Claim 43 has also been added, the text of which may be found in the body of the Amendment.

::ODMA\PCDOCS\DC2DOCS1\417406\1
VBHC Rev. 12/09/02 rpa